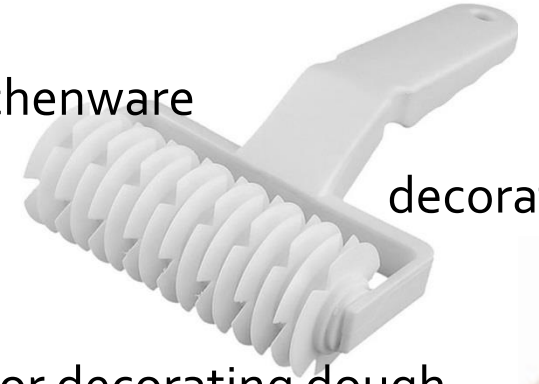
Converging evidence for the influence of semantic features on lexical diversity and geographical heterogeneity

Karlien Franco, Dirk Speelman & Dirk Geeraerts

pie cutter

baking tool

kitchenware



decorator

tool to flatten dough

tool for decorating dough

pizza roller cutter decorator

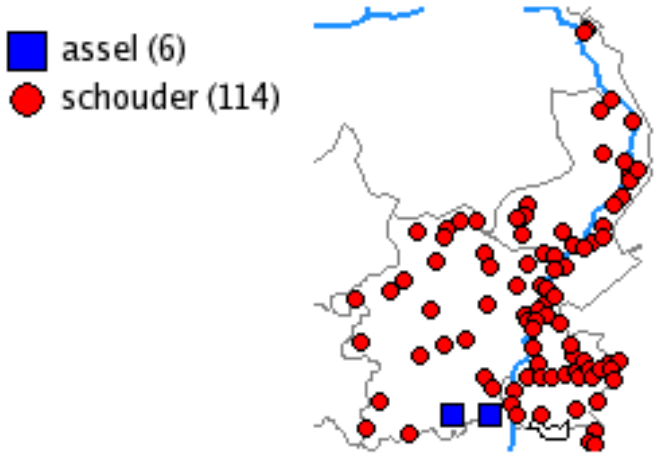
pie roller

cooking tool

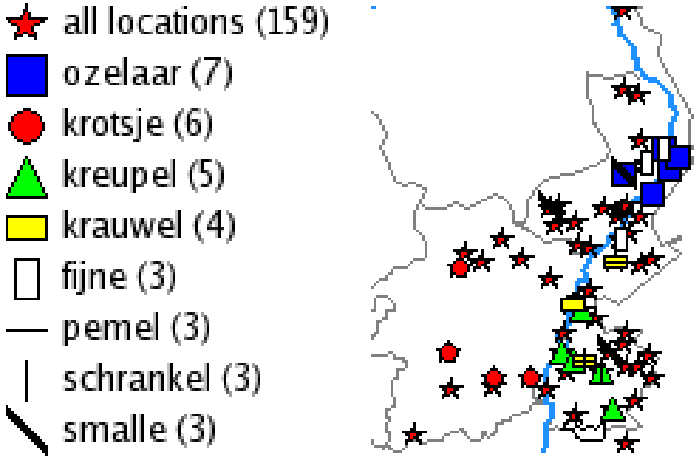
dialect data show dramatic variation

- in the amount of words that occur per concept
- in the degree to which these words are homogeneously spread across the geographical region where the dialect occurs

SHOULDER



VERY THIN AND WEAK PERSON



why do some concepts show more variation than others?

- concept characteristics influence the amount of lexical dialect variation in the semantic field of the human body

- more lexical geographical variability for concepts that
  - are prone to negative affect
  - have a low degree of onomasiological salience
  - are vague

(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)

negative affect

WELL BUILT WOMAN (GROF GEBOUWDE VROUW)	
machochel	mokkel
schommel	bai (fr.)
m	
m	
kapitein	dikke madam
mangel	dikke prij
machochel	flink wijf
schommel	fors vrouwmens
molenpaard	bammel
…	…

neutral

HEAD (HOOFD)
hoofd
kop

significantly more variation for concepts that are prone to negative affect

onomasiological salience

"various categories may have various degrees of entrenchment"

(Geeraerts, Grondelaers & Speelman 1999: 8)

onomasiological salience

not salient

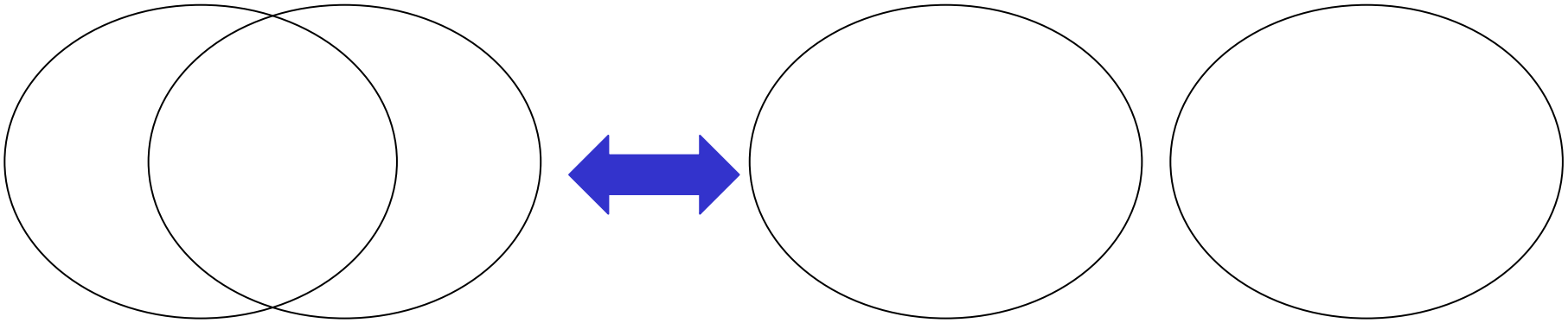
LITTLE DENTS BETWEEN THE KNUCKLES (KNOKKELKUILTJES)	
boelenhandjes	kuiltjes
deukjes	kussens
dompels	kwabbel
kinkd	
knob	
knokkelkuiltjes	putjes
knokkels	vetkuiltjes
knookjes	vingerkotjes
kotjes	vouwen
kreukeling	vouwtjes

highly salient

HEAD (HOOFD)
hoofd
kop

significantly more variation for concepts that are less salient

onomasiological vagueness



MODEST PEACEFUL, QUIET TUESDAY WEDNESDAY

vague

MODEST (INGETOGEN)		PEACEFUL, QUIET (KALM, BEDAARD)
bedaard	niet opvallend	bedaard
bedeesd	onopvallend	evenwichtig
bescheiden	op zijn eigen	gemoedelijk
charmant	ru	
deftig	ru	
eenvoudig	serieus	kalm
fatsoenlijk	simpel	koest
gemtlich (du.)	stemmig	ruhig (du.)
gewoon	stil	rustig
ingetogen	teruggetrokken	stil
kalm	zoet	traag
modest		zoet

not vague

TUESDAY (DINSDAG)	WEDNESDAY (WOENSDAG)
dinsdag	woensdag
	asgoensdag
	goensdag

significantly more variation for concepts that are vague towards neighbouring concepts

systematizing the results of the pilot studies

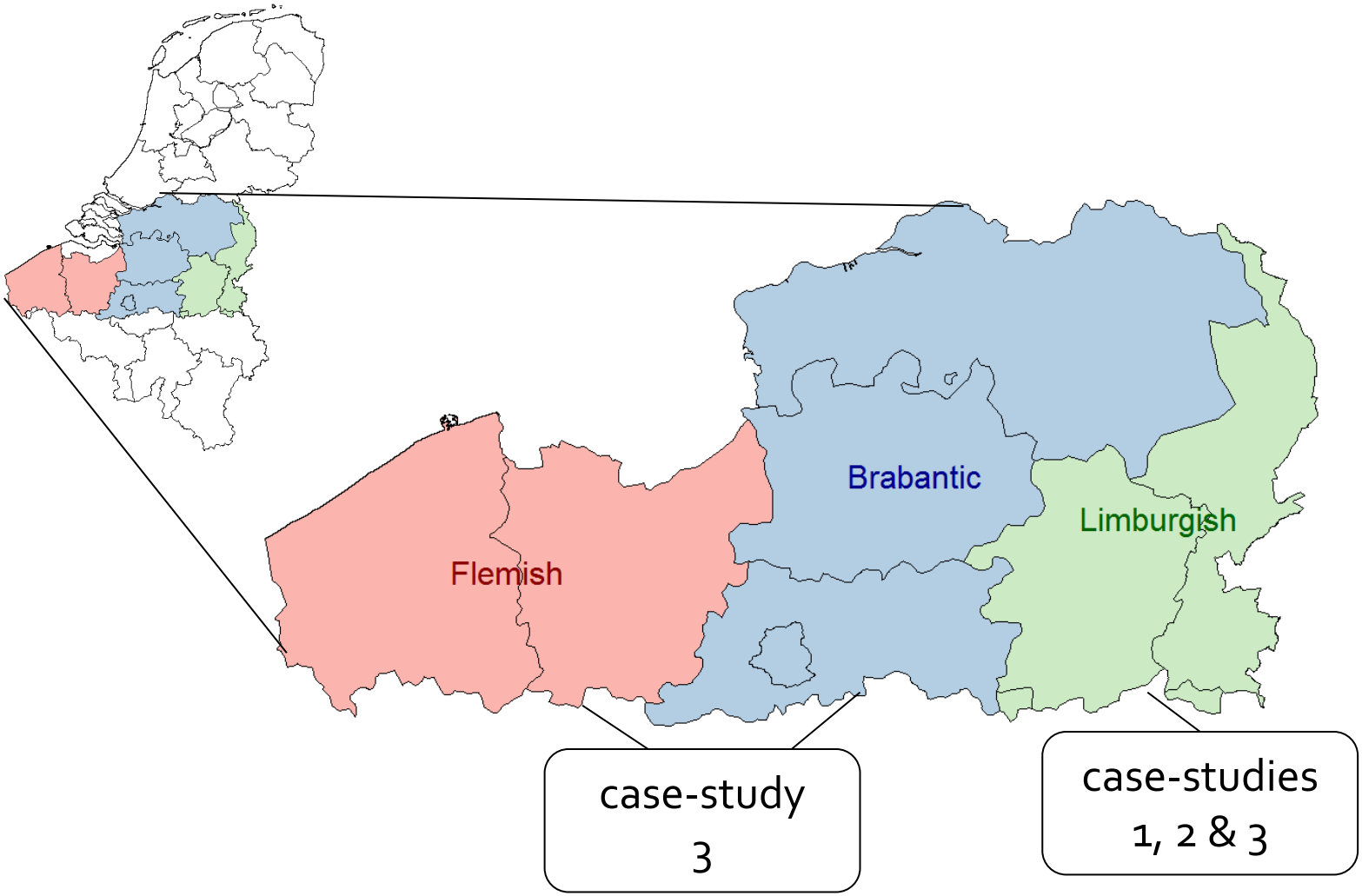
case-study 1
systematizing and extending on the pilot studies

case-study 2
concept features and the lexical component

case-study 3
zooming in on (experiential) salience

the dialects of Dutch

the dialects of Dutch



data

- databases of onomasiological dialect dictionaries
- thematically organized
- only data that was collected systematically with questionnaires

systematizing the results of the pilot studies

case-study 1
systematizing and extending on the pilot studies

case-study 2
concept features and the lexical component

case-study 3
zooming in on (experiential) salience

replication of pilot studies

SYSTEMATIZATION

- effect of concept characteristics in other fields than the human body
- measured per concept:
  - lack of salience
  - vagueness
  - affect

EXTENSION

- differences between semantic fields?
- measured at the level of the semantic field:
  - individual vs. community (e.g. Pickl 2013)
  - concrete vs. abstract concepts

design: six semantic fields

	concrete	**abstract**
man as an individual	the human body (4.390)	personality and feelings (2.347)
domestic life	the house (4.345)	family and sexuality (3.359)
community life	celebration and entertainment (3.772)	society, school and education (3.260)

design: six semantic fields

	concrete	**abstract**
man as an individual	the human body (4.390)	personality and feelings (2.347)
domestic life	the house (4.345)	family and sexuality (3.359)
community life	celebration and entertainment (3.772)	society, school and education (3.260)

(mean concreteness: Brysbaert et al. 2014)

methodology

RESPONSE VARIABLE: LEXICAL DIALECTAL VARIATION

- lexical diversity

  → some concepts have more different dialectal variants than others

- geographical fragmentation
  - dialect data is geographical in nature

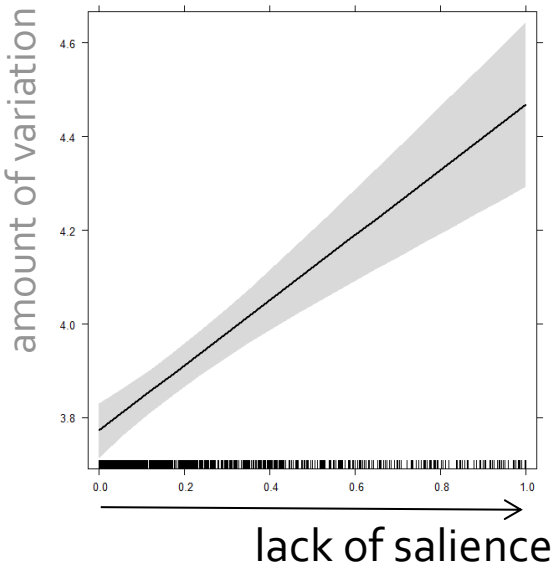
  → geographical scatter of variants can range from very homogeneous to very heterogeneous

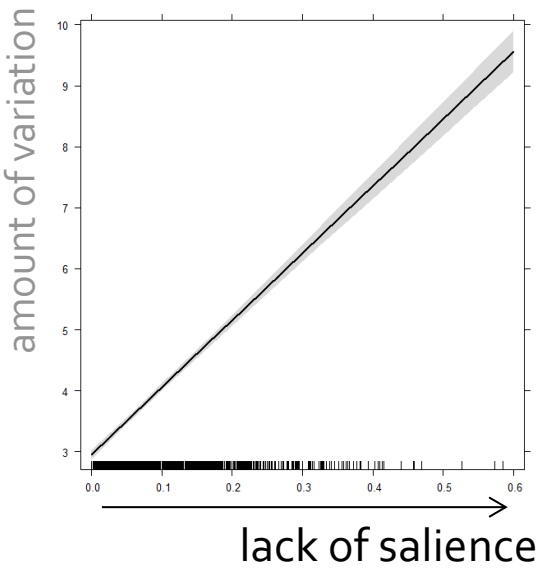
METHOD: LINEAR REGRESSION ANALYSIS

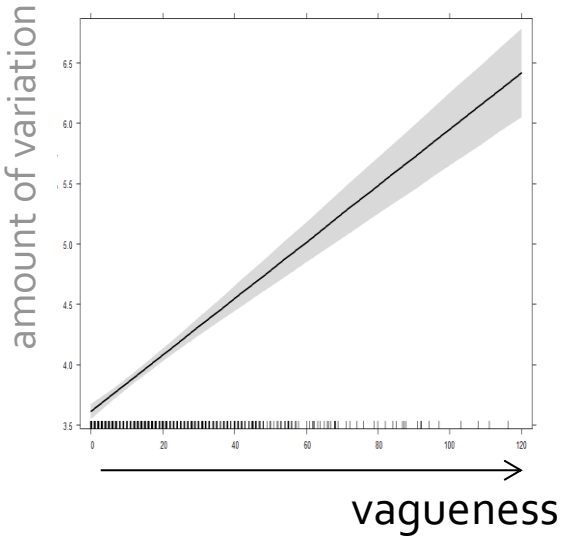
adjusted $R^2 = 0.6756$

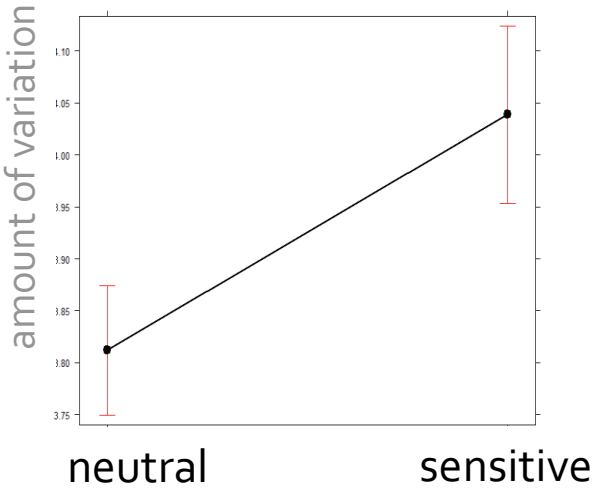
(Geeraerts & Speelman 2010, Speelman & Geeraerts 2008)

results: replication

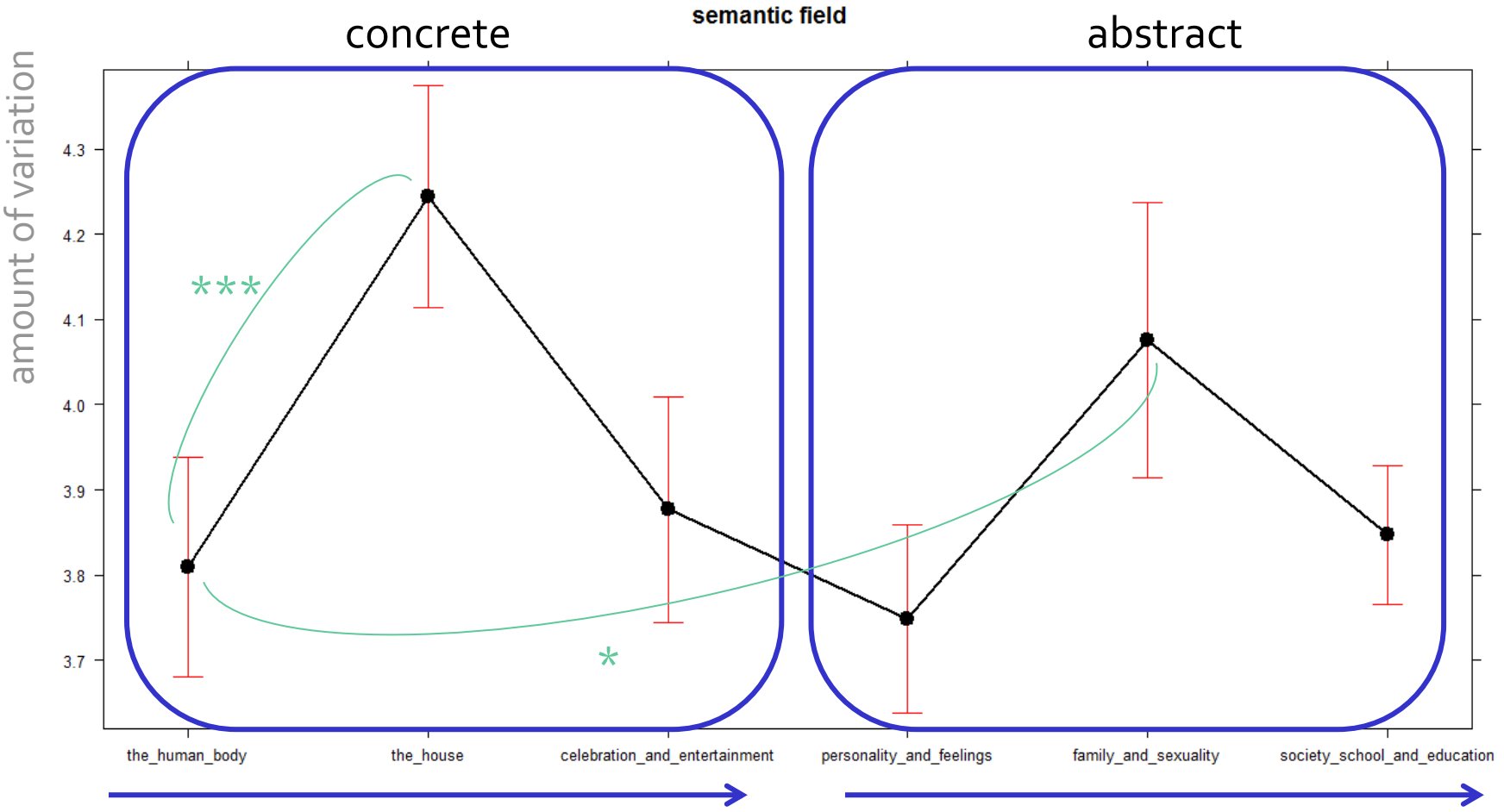








results: differences per semantic field?



local > society-related > universal?

discussion

SYSTEMATIZATION

lack of salience, vagueness and affect also lexical dialect variation in other fields than the human body

EXTENSION

- no clear effect of concreteness

  → on the concept-level?
- local > society-related > universal

systematizing the results of the pilot studies

case-study 1
systematizing and extending on the pilot studies

case-study 2
concept features and the lexical component

case-study 3
zooming in on (experiential) salience

research questions

do concept characteristics also influence variation in the lexicon-at-large?

two possible methodologies:

- data stratified along a different dimension than geography
- control for the geographical signal in dialect data

research questions

do concept characteristics also influence variation in the lexicon-at-large?

two possible methodologies:

- data stratified along a different dimension than geography
- control for the geographical signal in dialect data

  geography explains about 46% of the variability in purely lexical variation (lexical diversity)

  → are the results from case-study 1 still stable?

results

- model formula identical

- concept-related features all have significant, expected effect
  - more variation for less salient concepts
  - more variation for vaguer concepts
  - more variation for concepts prone to affect

- clear differences between semantic fields

  → some fields more prone to purely lexical variation

- less variation explained than in case-study 1 (adj. R^2 = 0.23 vs. 0.68)

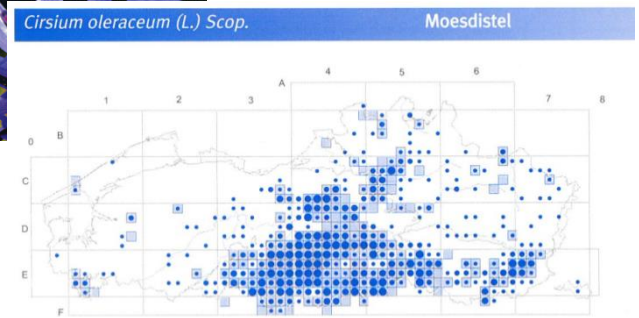
systematizing the results of the pilot studies

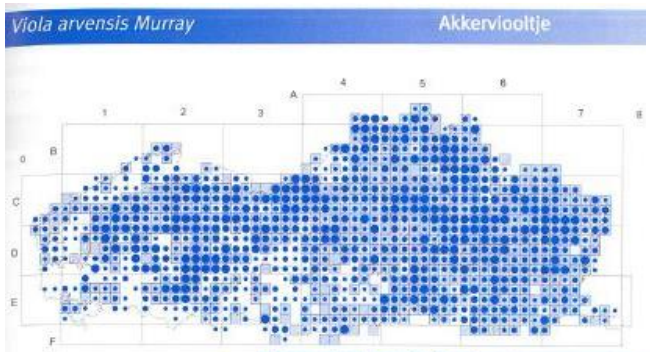
case-study 1
systematizing and extending on the pilot studies

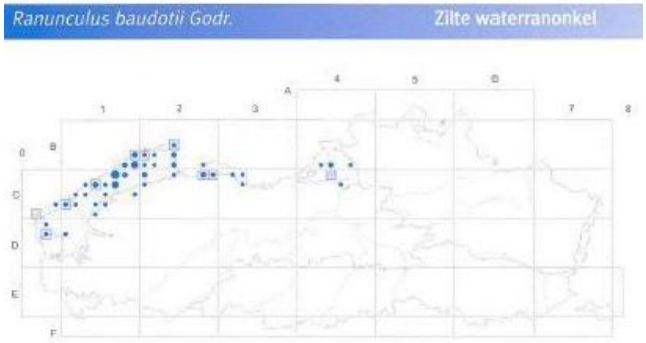
case-study 2
concept features and the lexical component

case-study 3
zooming in on (experiential) salience


ATLAS
VAN DE
FLORA
van VLAANDEREN en het BRUSSELS GEWEST


Cirsium oleraceum (L.) Scop.
Moesdistel


Viola arvensis Murray
Akkervlooltje


Ranunculus baudotii Godr.
Zilte waterranonkel

?

N = 137


WOORDENBOEK
VAN DE
LIMBURGSE DIALECTEN
Woordenboek van
de Vlaamse Dialecten
Woordenboek van de Brabantse Dialecten
Deel III
Sectie 4: De wereld tegenover de mens
Aflevering 3
Flora
Dr. J. Swanenberg
m.m.v. dr. H. Brok

methods & expectation

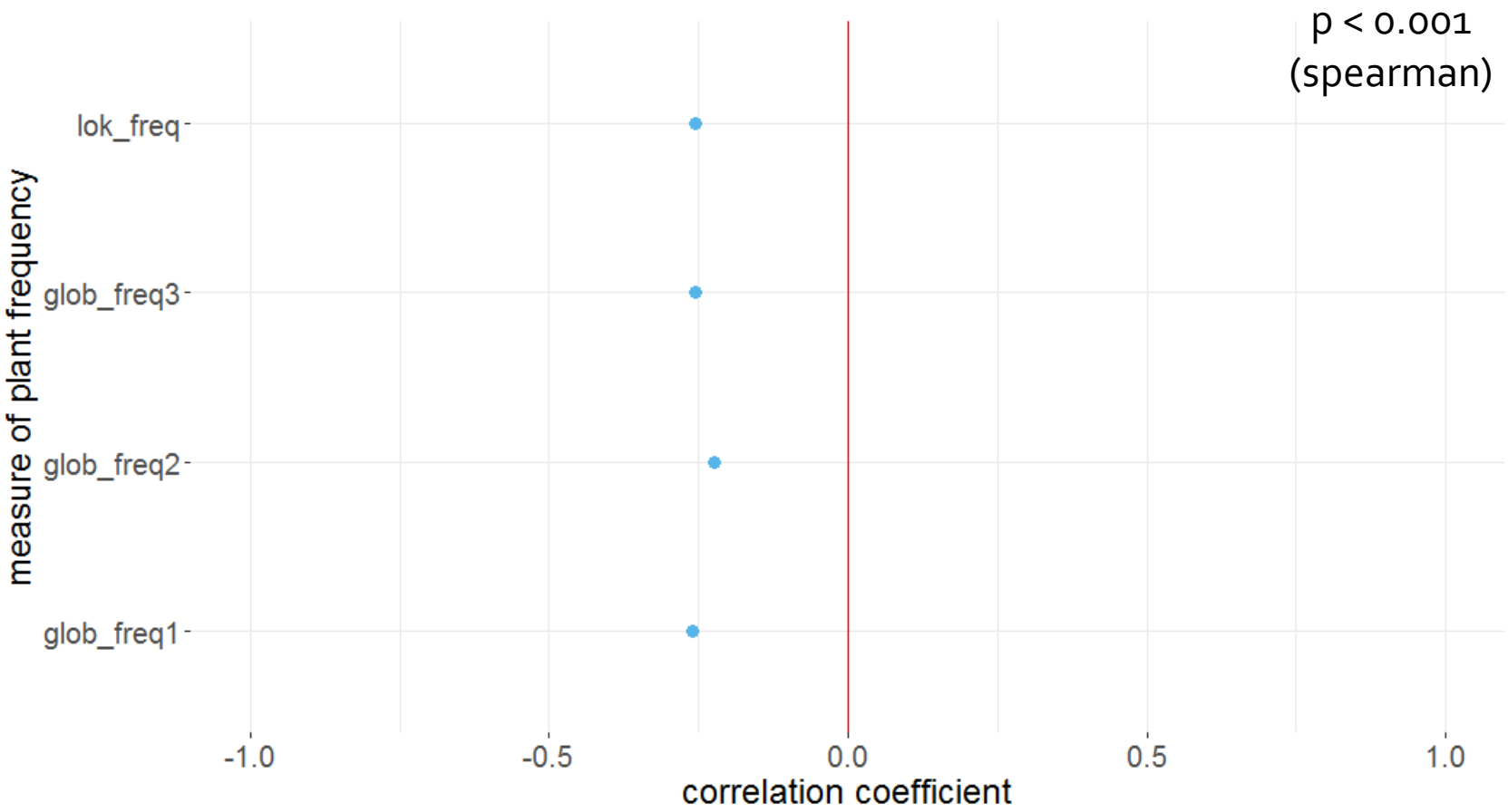
- amount of lexical variation per plant:
  - type-token ratio (TTR):

  = number of different lexemes (types) / number of records (tokens)
  - higher value = more variation
- experiential salience of a plant:
  - operationalized as plant frequency
  - 4 measures from Atlas, but highly correlated

- the relationship between plant frequency & lexical variation:
  - spearman rank correlation tests
  - correlation coefficients

  → negative correlations expected

results


p < 0.001
(spearman)
lok_freq
glob_freq3
glob_freq2
glob_freq1
measure of plant frequency
-1.0
-0.5
0.0
0.5
1.0
correlation coefficient

discussion

- significant negative correlation between plant frequency & lexical variation

  → less frequent plants show more lexical variation

- correlation coefficients low

  → other factors play a role as well

  - all plants relatively frequent
  - other measures of experiential salience?

    → significantly less variation for plants that…

    have a higher edibility rating

    have a higher medicinal rating

    → on average, poisonous plants also show less variation (NS)

conclusion: converging evidence

case-study 1	case-study 2	case-study 3
systematizing and extending on the pilot studies	concept features and the lexical component	zooming in on (experiential) salience
↓	↓	↓
concept features also affect variation in other semantic fields	concept features are relevant for variation in the lexicon-at-large	socio-cultural concept features correlate with the amount of lexical diversity

Thank you!

Questions? Suggestions?

karlien.franco
dirk.speelman
dirk.geeraerts
@ kuleuven.be

www.karlienfranco.com
http://wwwling.arts.kuleuven.be/qlvl/

concept-related predictors

1. LACK OF SALIENCE
   - proportion of missing places
     → ambiguous
   - proportion of multi-word expressions (MWE)
   - proportion of hapax legomena
   - prevalence (Keuleers et al. 2015)
     - word-level
     - missing data
2. VAGUENESS
   - number of types also used for other concepts (GS10, SG08)
3. AFFECT
   - manual, but relatively stable
   - mean valence (Moors et al. 2013), but missing data

from questionnaire …

KEUKEN-EN KOOKGEREI 2.

1. Hoe noemt u in uw dialekt:
   Het metalen voorwerp met hengsel en tuit dat dient om water in te koken?

   a. het moderne voorwerp (zie ill.) .....

2. b. het oudere voorwerp, dikwijls met ronde bodem, dat boven een haardvuur werd gehangen of in de opening van een kachel werd gezet (zie ill.)
   .....

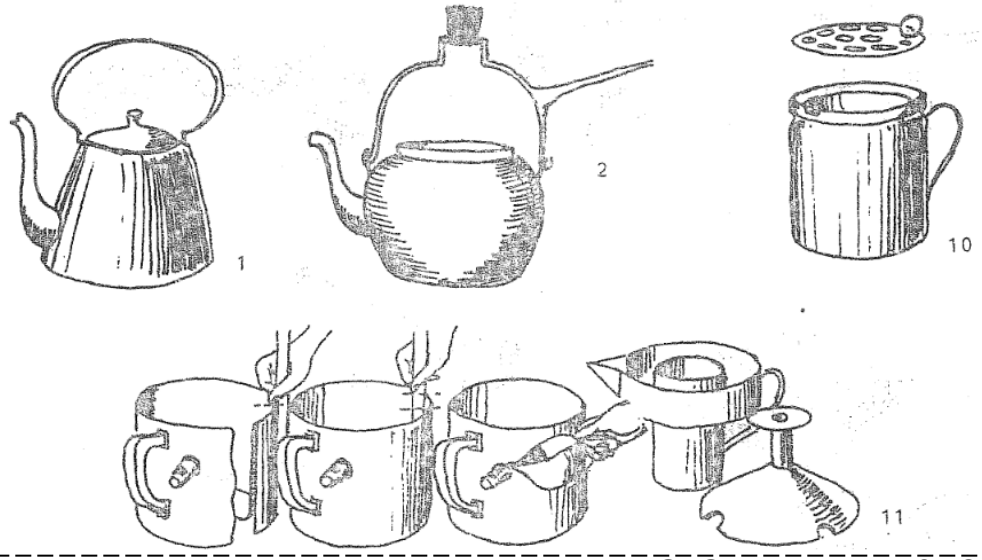
3. De pot met deksel, tuit en oor waarin koffie wordt gezet?
   .....

4. De zak waardoor het water over de gemalen koffie wordt gegoten?
   .....

5. De buikige pot met tuit en oor waarin enkel gezet?
   .....

6. Wanneer werd er thee gedronken? Als dagelijk als geneeskrachtige drank?

WOORDENBOEK VAN DE VLAAMSE DIALEKTEN : bijlage bij WVD 43 : huisraad
(Gelieve de illustraties niet terug te sturen)
Het nummer van de illustratie komt overeen met het nummer van de vraag.



… to dataset …

concept	variant	question	location	…
damesmantel ‘coat for women’	caban (fr.)	damesmantel, inventarisatie uitdrukkingen	Tervuren	…
overjas ‘overcoat’	frak	een jas die men over het colbert heen draagt	Leopoldsburg	…
…	…		…	…
vrolijk ‘cheerful’	spass (du.) haan	een opgeruimde, lichte, blijde stemming […]	Simpelveld	…
vrolijk ‘cheerful’	opgewekt	een opgeruimde, lichte, blijde stemming […]	Venlo	…
…	…		…	…

… to measurements at the level of the concept

concept	**lexical geographical variation**	**predictor 1: affect sensitivity**	**predictor 2: vagueness**	**…**
achterdochtig 'suspicious'	5	sensitive	2.275	
achterhoofd 'back of the head'	21	neutral	4.977	…
…	…	…		…
speelplaats 'playground'	3	neutral	2.341	…
speels 'light-hearted'	9	sensitive	3.561	…
…	…	…		…
…	…	…		…